An investigation on the use of iVectors for Improved ASR Robustness

Dimitrios Dimitriadis, Samuel Thomas

Department of Advanced LVCSR
Watson, IBM, Yorktown Heights, US
[dbdimitr, sthomas]@us.ibm.com

Sriram Ganapathy

Department of EE
Indian Institute of Science, Bangalore, India
sriram@ee.iisc.ernet.in

Abstract

In this paper we propose two different ivector representations that improve noise robustness of automatic speech recognition (ASR). While the first kind of ivectors is derived from ”noise only” components of speech as provided by an adaptive MMSE denoising algorithm, the second variant is extracted from mel filterbank energies containing both speech and noise. The effectiveness of both these representations is shown by combining them with two different kinds of spectral features - the commonly used log-mel filterbank energies and Teager energy cepstral coefficients (TECCs). Using two different DNN architectures for acoustic modeling - a standard state-of-the-art sigmoid-based DNN and an advanced architecture using leaky ReLUs, dropout and rescaling, we demonstrate the benefit of the proposed representations. On the Aurora-4 multi-condition training task the proposed front-end improves ASR performance by 4%.

Index Terms: speech recognition, noise robustness, feature extraction, ivectors

1. Introduction

2. Noise Signal Estimation

The training and testing sets of the Aspire task are heavily mismatched. Amongst the reasons for this mismatch is the existence of reverberation and additive noise in the testing set. For example, the SSNR in the dev and the dev-test sets are estimated around $13.2dB$ and $12.8dB$ correspondingly. In order to decrease the effects of these two types of noise on the ASR performance, we first suppress the additive noise using a variation of the MMSE algorithm [1]. Then, we subtract the late reverberation component of the signal employing the MSLP (”long-term Multi-Step Linear Prediction”) algorithm [2]. The parameters for both algorithms are fine-tuned on the dev test, yielding an absolute improvement in ASR performance of 4%.

In more detail, the denoising process is applied only on the dev and dev-test sets, while the audio of the training set is left unprocessed. The denoising system consists of a two-step approach: the first N frames (fine-tuned to 10 frames here) of speech are used to evaluate whether the SSNR is low enough to run the denoiser or keep the audio file unprocessed, then an MMSE-based noise suppresion system is employed. The denoising algorithm is based on the minimum mean-square error (mmse) estimation of the noise power. The speech component of noisy audio is then obtained by multiplying the noisy power spectrum by a gain [3]. However, this process suffers from leakage of the speech power to the noise estimates. In order to minimize it, a time- and frequency-dependent smoothing parameter is proposed in [1], where the estimate of speech presence probability is also investigated. Further, the gain function is trained by an iterative data-driven training method [4] and look-up table is created based on the speech and noise variance estimates. A safety net is also employed for the cases when the noise levels suddenly increase, as described in [1].

3. Feature Extraction

3.1. Auditory Filterbank

The notion of the Equivalent Rectangular Bandwidth (ERB) is introduced to quantify, in a way, the bandwidth of asymmetrical filters like the auditory ones. More specifically, given the $|H(f)|$ as the magnitude of the filter frequency response and $|H(f_{max})|$ it’s maximum gain obtained in f_{max}, then the filter *ERB* (in Hz) is defined by

$$ERB = \frac{\int |H(f)|^2 df}{|H(f_{max})|^2} \tag{1}$$

The ERB is the equivalent bandwidth of an orthogonal filter with constant gain $|H(f_{max})|$ when its energy (the integral of its frequency response) equals to the corresponding one of the integral of $|H(f)|^2$.

Recent studies of the human hearing physiology [5, 6, 7] have shown that the human physiology dictates that the auditory filter bandwidths are given by the $ERB(f)$ function

$$ERB(f) = 6.23(f/1000)^2 + 93.39(f/1000) + 28.52 \tag{2}$$

where f is the filter center frequency in Hz. Moreover, the filter placing is equidistant in the *Critical* (bark) frequency scale

$$bark(f) = \frac{26.81f}{f+3920} - 0.53 \tag{3}$$

where $0 \leq f \leq F_s/2$. Finally, a good approximation of the auditory filters are the asymmetrical Gammatone filters

$$g(t) = At^{n-1}\exp\left(-2\pi bERB(f_c)t\right)\cos(2\pi f_c t) \tag{4}$$

where A, b, n are the Gammatone filter design parameters and f_c its center frequency. In [5] it is proposed that the auditory filters should have $b = 1.019$ and $n = 4$. Thus, the filter frequency response $G(\omega)$ is given by

$$\begin{aligned} G(\omega) &= \frac{A}{2}\frac{6}{\left(2\pi bERB(f_c) + j(\omega - \omega_c)\right)^4} + \\ &+ \frac{A}{2}\frac{6}{\left(2\pi bERB(f_c) + j(\omega + \omega_c)\right)^4} \end{aligned} \tag{5}$$

Moreover, the filter gain A is set taking under consideration that $|H(\omega_c)| = 1$ and

$$A = \frac{1}{\sum_{k=1}^{N} t^{n-1}\exp\left(-2\pi bERB(f_c)t\right)} \tag{6}$$

where N the length of the discretized impulse response in samples.

The main differences between the proposed filterbank and the typical one used for MFCC estimation are the type of filters used and their bandwidths. The auditory filterbank proposed is not costant-Q and emphasizes the lower part of the frequencies where the main part of the acoustic information is located. The recognition results show that the need for constant-Q filterbanks is not always the appropriate solution and mimicking the human auditory system could turn out to be a better approach providing robustness to noise and improved results.

A Gammatone filterbank, with filters placed according to the bark scale and bandwidths given by the $ERB(f)$ is very close to the human auditory system [8, 9, 6]. The human ear employs several thousand of filters and the corresponding filterbank is very dense. On the contrary, following the $ERB(f)$ 'constraint', when the number of filters is small (in the range of 30 filters), the yielded filterbank is quite sparse. For this reason, we introduce a multiplying factor F to the filter bandwidth curve, re-adjusting it. Experimental results [10], provide clear indication that this factor, and thus the filter bandwidths, are important to the recognition process. This parameter regulates the filter overlap, and after experimentation, we have concluded that the optimal values range in $1.0 - 2.0$ depending on the signals' SNR. Here, we present a 25-filter Gammatone filterbank with *ERB* factor $F = 1.5$. The filter placing is according to the Bark frequency scale.

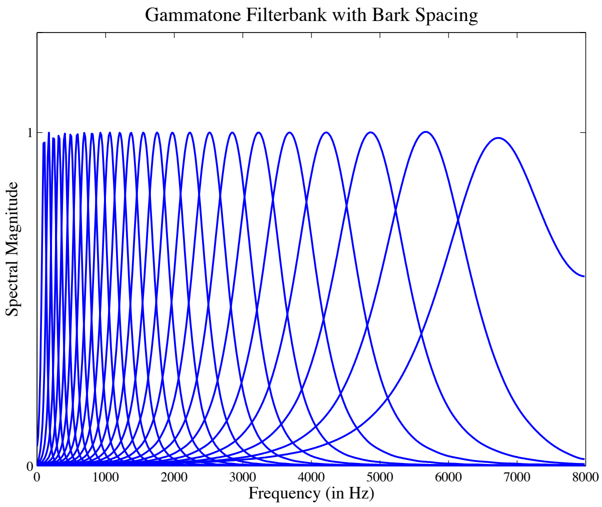

Figure 1: *25-filter Bark-spaced Gammatone filterbank with ERB factor* $F = 1.5$.

The typical logmel are estimated over a filterbank of triangular filters with 50% overlap as the log mean squared amplitudes of the bandpass signals, [11]. On the other hand, we propose incorporating information about the time-varying nature of speech using the instantaneous Teager-Kaiser (TK) energy instead of the typical approach. This way, the acoustic information the features' is 'richer'. In addition, we use an auditory-inspired filterbank, [12], instead of the triangular filterbank taking advantage of the human hearing process. The proposed features are shown to be more robust in additive noise and provide additional acoustic information when compared to the logmels. These auditory filters are implemented by Gammatone filters and they are smoother and broader than the triangular filters.

The *TEC* estimation algorithm is described with the following steps:

i. Use a Gammatone filterbank to estimate a sequence of bandpass, speech signals. The number of filters is ranging from 25 to 200 filters,

ii. Estimate the mean TK-energy for each one of the framed bandpass signals,

iii. Estimate the Spectral coefficients as the log mean energies , and

The first two steps combine the auditory filtering scheme with the more 'natural' approach of the speech TK-energy notion. These steps differentiate the proposed algorithm from the typical MFCC extraction algorithm. The ASR results show significant improvement, especially in noisy recognition tasks [12].

4. IVector Extraction

5. Experiments

The proposed techniques are evaluated using a series of experiments on Aurora 4 - a medium vocabulary task, based on the Wall Street Journal corpus [13]. Using the task's multi-condition training set with 7137 utterances sampled at 16kHz from 83 speakers, neural network based acoustic models are first trained and then tested on a set of 330 utterances from 8 speakers. While one half of the training utterances is recorded with a primary Sennheiser microphone, the second half is collected using one of 18 other secondary microphones. Both halves have clean and noisy speech utterances. The noisy utterances are corrupted with one of six different noise types (airport, babble, car, restaurant, street traffic and train station) at 10-20 db SNR.

Similar to the train set, the test sets are also recorded over multiple microphones - a primary microphone and one other secondary microphone. In addition to clean test data collected over each of these microphones, the same six noise types used in train are employed to create noisy test sets at 5-15dB SNR, resulting in a total of 14 test sets. These test sets are commonly grouped into 4 subsets - clean (test set A), noisy (test set B), clean with channel distortion (test set C) and noisy with channel distortion (test set D).

Before building deep neural network (DNN) baselines for multi-condition training, an initial set of HMM-GMM models are trained to produce alignments for the multi-condition training utterances. Unlike the baseline systems, these models are built on the corresponding clean training (7137 utterances) set of the Aurora 4 task in a speaker dependent fashion. Starting with 39-dimentional VTL-warped PLP features and speaker based cepstral mean/variance normalization, an ML system with FMLLR based speaker adaptation and 2000 context-dependent HMM states is trained. The alignments produced by this system, are further refined using a DNN system also trained on the clean training set with FMLLR based features.

For both these architectures separate systems using sigmoid and leaky ReLU non-linearities are trained for comparisons. All the systems are trained on 40 dimensional *log-mel* and *TEC* spectra augmented with Δ and $\Delta\Delta$s. The *log-mel* spectra are extracted by first applying *mel* scale integrators on power spectral estimates in short analysis windows (25 ms) of the signal followed by the *log* transform. Each frame of speech is also appended with a context of 11 frames after applying a speaker independent global mean and variance normalization.

The DNN systems estimate posterior probabilities of 2000 output targets using a network with either 6 or 7 hidden layers, each having 1024-2048 units per layer. For the DNN systems using leaky ReLU non-linearities, we additionally use a fixed dropout of 50%. Instead of applying dropouts on all layers, our best results are obtained when dropouts are selectively applied only on the third and fourth hidden layers, only when the pre-training of the networks is finished. Similarly, we have also applied a fixed drop-in rate of 20% for the input features. Finally, rescaling of the weights is performed after every mini-batch iteration. All DNNs are discriminatively pre-trained before being fully trained to convergence. After training, the DNN models are decoded with the task-standard WSJ0 bigram language model.

Table 1: *Aurora-4: Multi-condition Training: Sigmoid 6h x 1024.*

Multi-condition Training: Sigmoid					
	A	B	C	D	Aver.
logmel	6.31	11.98	13.06	24.96	17.21
TEC	6.84	11.52	13.30	24.07	16.69
logmel+Noise iVecs					
logmel+Noisy iVecs					
logmel+Noise+Noisy iVecs	6.09	10.66	12.96	22.75	15.68
TEC+Noise iVecs					
TEC+Noisy iVecs	6.93	11.59	13.75	23.27	16.41
TEC+Noise+Noisy iVecs					

Table 2: *Aurora-4: Multi-condition Training: Sigmoid 7h x 2048.*

Multi-condition Training: Sigmoid					
	A	B	C	D	Aver.
logmel	5.85	10.33	11.13	22.64	15.34
TEC	6.07	10.24	11.34	21.82	14.98
logmel+Noise iVecs	5.34	9.81	11.54	21.13	14.47
logmel+Noisy iVecs	5.51	10.43	11.41	22.44	15.29
logmel+Noise+Noisy iVecs	5.38	9.79	11.58	21.74	14.72
TEC+Noise iVecs	6.00	10.43	11.86	22.38	15.34
TEC+Noisy iVecs	5.70	9.70	12.27	21.38	14.61
TEC+Noise+Noisy iVecs	5.62	9.83	12.54	21.22	14.60

6. Discussion

Under both language models, both CNN and DNN systems built with ReLU non-linearities perform considerably better than corresponding systems with sigmoid non-linearities. The CNN based systems are also always better than DNN based systems although the input features are identical and the number of parameters comparable.

Table 3: *Aurora-4: Multi-condition Training: ReLU 6h x 1024.*

Multi-condition Training: ReLU					
	A	B	C	D	Aver.
logmel	4.13	7.46	7.34	16.19	10.96
TEC	4.50	7.32	7.58	15.48	10.64
logmel+Noise iVecs	4.28	7.32	7.75	16.78	11.19
logmel+Noisy iVecs	4.35	7.35	8.14	16.73	11.21
logmel+Noise+Noisy iVecs	4.00	7.17	7.70	16.41	10.94
TEC+Noise iVecs	4.24	7.37	7.62	16.26	10.98
TEC+Noisy iVecs	4.75	7.11	7.81	15.41	10.55
TEC+Noise+Noisy iVecs	4.26	7.12	7.64	15.44	10.51

Table 4: *Aurora-4: Multi-condition Training: ReLU 7h x 2048.*

Multi-condition Training: ReLU					
	A	B	C	D	Aver.
logmel	4.30	7.69	7.94	16.86	11.39
TEC	4.60	7.45	7.85	16.82	11.29
logmel+Noise iVecs					
logmel+Noisy iVecs					
logmel+Noise+Noisy iVecs					
TEC+Noise iVecs					
TEC+Noisy iVecs					
TEC+Noise+Noisy iVecs	4.45	7.54	8.11	16.65	11.27

7. Conclusions

Authors must proofread their PDF file prior to submission to ensure it is correct. Authors should not rely on proofreading the Word file. Please proofread the PDF file before it is submitted.

8. Acknowledgements

The ISCA Board would like to thank the organizing committees of the past INTERSPEECH conferences for their help and for kindly providing the template files.

9. References

[1] J. S. Erkelens and R. Heusdens, “Tracking of nonstationary noise based on data-driven recursive noise power estimation,” *IEEE Trans. on Audio, Speech and Language Process.*, vol. 16, no. 6, pp. 1112–1123, Aug. 2008.

[2] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, “Suppression of Late Reverberation Effect on Speech Signal Using Long-Term Multiple-step Linear Prediction,” *IEEE Trans. on Audio, Speech and Language Process.*, vol. 17, no. 4, May 2000.

[3] Y. Ephraim and D. Malah, “Speech enhancement using a minimum min-square error log-spectral amplitude estimator,” *IEEE Trans. on Acoust., Speech and Signal Process.*, vol. 33, no. 2, pp. 443–445, 1985.

[4] J. S. Erkelens, J. Jensen, and R. Heusdens, “A data-driven approach to optimizing spectral speech enhancement methods for various error criteria,” *Speech Communication*, vol. 49, pp. 530–541, Aug. 2007.

[5] T. Irino and R. D. Patterson, “A Time-Domain, Level-Dependent Auditory Filter: The Gammachirp,” *Journ. Acoustical Society of America*, 1997.

[6] O. Ghitza, “Auditory Models and Human Performance in Tasks Related to Speech Coding and Speech Recognition,” *IEEE Trans. Speech and Audio Processing*, 1994.

[7] B. R. Glasberg and B. C. J. Moore, “Derivation of Auditory Filter Shapes from Notched-Noise Data,” *”Hear. Res.”*, 1990.

[8] H. Hermansky and N. Morgan, “Rasta processing of speech,” *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, Aug. 1994.

[9] H. Hermansky, “Perceptual linear predictive (plp) analysis of speech,” *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–1752, Aug. 1990.

[10] D. Dimitriadis, P. Maragos, and A. Potamianos, “On the effects of filterbank design and energy computation on robust speech recognition,” *IEEE Trans. on Audio, Speech and Language Process.*, vol. 19, no. 6, pp. 1504–1516, Aug. 2011.

[11] S. B. Davis and P. Mermelstein, “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences,” *IEEE Trans. Acoust., Speech, Signal Processing*, 1980.

[12] D. Dimitriadis, P. Maragos, and A. Potamianos, “Auditory Teager Energy Cepstrum Coefficients for Robust Speech Recognition,” in *Eurospeech*, 2005.

[13] N. Parihar and J. Picone, “Aurora Working Group: DSP Front-end and LVCSR Evaluation AU/384/02,” Inst. for Signal and Information Processing, Mississippi State University, Tech. Rep., 2002.